水平社交能力的社交机器人，要么从小数据出发获得能够保证隐私安全但社交能力显得落后或匮乏的社交机器人。前者不仅具有高水平的社交能力，在其他知识领域中同样表现卓越，但同时保持了大数据的所有风险。而后者也不失为一个选择，就像不掌握人类语言的动物也能起到很好的陪伴作用一样，对不同的人来说，他们并不都需要或期待陪伴者拥有高水平社交能力。对这二者的选择实际上是基于不同价值的选择，前者更注重社交能力的发展，后者更注重安全性。因此，在社交机器人技术的发展中，小数据主义可以在一定领域中和需要下发挥优势，这需要人类对社交机器人抱有不同的期待。

机器人惩罚缺漏争论及其拟人化分析*

孙　圣

惩罚缺漏（retribution gaps）由约翰·丹纳赫（John Danaher）通过简单的思想实验提出。惩罚缺漏主要指在与机器人有关的伤害案件中，因无法确定报复性指责的对象而无法找到合适的惩罚对象，从而出现伦理和司法意义上的缺漏。丹纳赫认为，惩罚缺漏是机器人具有自主能力的必然结果，机器人和生产商都不是适当的惩罚对象。①

惩罚缺漏常被用作处理自动驾驶汽车和养老机器人等伤人案件中的伦理问题的判决标准。② 此外，还有很多与惩罚缺漏相关的前沿争论，如能动性

* 本文系甘肃省科技计划项目（基础研究计划）软科学专项“甘肃省科技创新与思想实验交叉学科建设路径研究”（23JRZA398）、国家社会科学基金西部项目“模态反驳模式下的思想实验可靠问题研究”（23XZX003）的阶段性成果。西北师范大学哲学与社会学院2024级硕士研究生岳恒山参与文献检索与整理，在此特表感谢。

① 参见John Danaher, Robots, Law and the Retribution Gap, *Ethics and Information Technology*, Vol. 18(4), 2016, pp. 299−309。

② 参见［瑞典］斯文·尼霍姆:《人与机器人：伦理、行动与拟人论》，刘铮译，上海交通大学出版社2024年版，第60~64页。

从人类行动者到机器人行动者的转移，① 以及如何建立“风险共同体”以避免“道德运气”的影响。② 然而，将惩罚缺漏作为判决标准颇具争议。

在拟人化（anthropomorphism）和透明性问题中，对将惩罚缺漏作为判决标准的反对声音受到较多关注。③ 对此，丹纳赫认为，在对惩罚缺漏的讨论中可悬搁拟人化问题。④ 实际上，丹纳赫的惩罚缺漏理论采用了静态视角。而静态视角是逻辑实证主义具有的常见隐患，易引发偏见，且难以察觉。⑤

在笔者看来，丹纳赫对反对将惩罚缺漏作为判决标准的声音尤其是拟人化问题没有给予足够重视。本文将分别从惩罚是否适度、惩罚的合理对象是谁、拟人化的客观性以及拟人度的技术手段四个角度展开论述。

一、惩罚缺漏还是惩罚过度？

实际上，考察与技术前景有关的议题，不能完全脱离其哲学前提，对元宇宙在未来不是什么的争论就是一个例子。⑥“拟人化”一词源于大卫·休谟，他认为人类有将所有生物设想为与人类相像的倾向。⑦ 在人工智能领域，拟人化则是一种试图将机器人等非人类实体与人类特征建立联系的手段，这些

① 参见 Mark Coeckelbergh, Responsibility and the Moral Phenomenology of Using Self−driving Cars, *Applied Artificial Intelligence*, Vol.30(8), 2016, pp. 748−757。

② 参见 Alexander Hevelke and Julian Nida−Rümelin, Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis, *Science and Engineering Ethics*, Vol.21(3), 2015, pp. 619−630。

③ 参见 Taemie Kim and Pamela Hinds, Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human−robot Interaction, *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, New York: IEEE, 2006, pp. 80−85。

④ 参见John Danaher, Robots, Law and the Retribution Gap, *Ethics and Information Technology*, Vol.18(4), 2016, pp. 299−309。

⑤ 参见孙圣：《思想实验复杂性及其多时间尺度出路——兼论费米悖论》，《系统科学学报》2023 年第 4 期，第 23~28、34 页。

⑥ 参见孙圣、赵月刚：《元宇宙是所有可能世界吗？——从思想实验与时间真实性出发的知识论研究》，《浙江社会科学》2023 年第 4 期，第 110~118 页。

⑦ 参见 David Hume, *The Natural History of Religion*, London: A. and H. Bradlaugh Bonner, 1889, p. 11。

特征包括意图、动机、人类情感和行为等。[①] 而亚里士多德意义上的道德判断取决于人是否以符合德行的方式行事。

这样的思想渊源导致相关领域的学者与企业家普遍认为，对机器人的道德判断与是否为机器人赋予人类特征是直接相关的，其中的代表人物就包括人们熟悉的埃隆·里夫·马斯克（Elon Reeve Musk）。由此，一个疑问进入了学界的视野：离开拟人化的机器人能否作为惩罚的合理对象？

从丹纳赫本人的论述中不难发现，他之所以提及惩罚缺漏，是因为自主能力较强的机器人施加的侵害行为使人类难以通过本能或直觉确认报复的对象。[②] 我们知道，机器人越不像人，就越会被认为不用为事故负责。[③] 但如果机器人能够实现一定程度的拟人化，则断言的前提是存疑的。

目前，“智慧”和“富有同情心”等拟人化标签已被认为是决定机器人销量的重要因素，从而引发机器人设计上的改变。特斯拉电动车采用拟人化的视觉识别系统而非雷达测距系统，其原因除了成本控制的要求，也不排除这一市场因素。而一旦将拟人化的机器人作为道德主体，人类的报复欲望就有了输出的对象。在较长时间尺度上看，这一情景将逐步成为现实——只要能激起人的报复欲望，就能够实现惩罚对象的确认。因此，不会出现丹纳赫担心的惩罚缺漏。

然而，对机器人的“惩罚”还可能源于人类报复欲望之外的其他类型的心理活动。即便不是出于丹纳赫强调的报复心理，而是出于其他心理，人们也可能对机器人施加广义的“惩罚”。这是因为，在法律能够规定的所有惩罚类型之外，人们还会自发地产生对机器人的“惩罚”行为。例如，人类个体会将负面情绪无意识地转移到比自己弱小或地位更低的对象上，这在心理学和社会学上叫做“踢猫效应”（kick the cat effect）。[④] 由此产生的私下的“惩罚”不仅不会导致丹纳赫意义上的惩罚缺漏，甚至还会导致惩罚过度。

---

① 参见 Adam Waytz，Nicholas Epley and John T. Cacioppo, Social Cognition Unbound: Insights into Anthropomorphism and Dehumanization, *Current Directions in Psychological Science*, Vol.19(1), 2010, p. 59。

② 参见 John Danaher, Robots, Law and the Retribution Gap, *Ethics and Information Technology*, Vol.18(4), 2016, pp. 299–309。

③ 参见 Taemie Kim and Pamela Hinds, Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human–robot Interaction, *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, New York: IEEE, 2006, pp. 81, 84。

④ 参见张文成：《墨菲定律》，古吴轩出版社 2017 年版，第 40 页。

在惩罚缺漏的心理学来源上，丹纳赫的论述存在缺陷，这会导致一系列的应用障碍。人们对机器人做出的源于“踢猫效应”的“惩罚”行为，并非出于丹纳赫认为的报复心理，而是出于“欺负心理”。源于“踢猫效应”的对机器人的“惩罚”，不仅是对机器人的惩罚过度，如不加以限制，还会带来对施加欺负行为的人类的惩罚缺漏。

二、关于惩罚缺漏的对象是否存在争议？

“惩罚”一词连接的对象通常是未能尽到在法律、道德规范、伦理范畴内的基本义务或触犯了禁止性规定的人。然而，当机器人具有某种程度的自我意识时，从机器人的视角看，它能否根据相应的法律、道德规范、伦理要求规划其自身的行为？这种规划行为与人类科学家进行的具身性的（embodied）思想实验 ① 具有某种相似性。需要具身地考虑机器人的处境，否则势必会导致新的混乱。

过度关注对机器人的惩罚而忽略对真正施加侵害的人类的惩罚，会产生新的惩罚缺漏问题。在争论对哪个对象存在惩罚上的缺漏时，如果有意绕开机器人背后的人类生产商，就会导致对拟人化技术的滥用。在美国人工智能学家、企业家阿德里安娜·普拉卡尼（Adriana Placani）看来，机器人拟人化不仅会导致惩罚焦点向机器人转移，还易导致人类个体被错误地免责。 ②

从“人机共同体”概念出发的机器人相关立法对生产商积极性的保护 ③ 同样不应被过分强调。首先，对法律的经济分析被怀疑不过是一种特定的增强资本主义自由市场体系的意识形态偏好，对此的过度讨论会削弱社会的公平与正义。 ④ 其次，从长远利益上看，不强调“人机共同体”反而有利于保护生产商的积极性，因为还应考虑市场需求等因素。由“人机共同体”而导致的错误责任认定会让购买者和使用者承担不应有的连带赔偿责任。即便可

① 参见孙圣：《机会主义是由具身认知向度导致的吗？——对等效原理若干科学史争论的新审视》，《自然辩证法研究》2024 年第 4 期，第 115~122 页。

② 参见 Adriana Placani, Anthropomorphism in AI: Hype and Fallacy, *AI and Ethics*, Vol.4, 2024, p. 696。

③ 参见 Matthew R. Gaske, Artificial Intelligence Regulation, Minimum Viable Products, and Partitive Innovation, *Emory Law Journal Online*, Vol.73, 2023, p. 25。

④ 参见［英］瓦克斯：《法哲学：价值与事实》，谭宇生译，译林出版社 2013 年版，第 67 页。

以通过后续的补救途径进行追偿，这也在无形中增加了购买者和使用者的成本，降低了其购买和使用机器人产品的意愿，从而减少了机器人产品的市场份额，间接地打击了生产商的积极性。

为了更深刻地说明惩罚缺漏的错误对象指向包含理论隐患，我们以一个思想实验为例，说明当自动驾驶机器人面对类似于“电车问题”的伦理选择时需要考虑对人的惩罚，否则责任的归属问题与惩罚对象的确认问题将是无解的。考虑一个情形，前方道路狭窄，汽车行驶速度较快，此时突然有行人横穿马路，面对这一情形自动驾驶机器人仅有两个选项：刹车后撞上行人，或猛打方向盘撞上路边的障碍物并导致车内人类受伤。

无论如何选择，显然自动驾驶机器人将做出违背机器人三定律的伤人行为。机器人三定律由于在很多特定情景中存在不自洽风险，难以被单独作为实现机器人自我规划的原则基础。机器人三定律的存在本身就是一种对机器人的惩罚过度。在行人家属看来，机器人三定律自身的缺陷间接地造成了事故的惩罚缺漏。惩罚缺漏是存在的，其理论意义重大，可作为对机器人三定律这种基础假设的判决标准。但是，在上述思想实验中，惩罚缺漏存在的问题不是丹纳赫意义上的。

因此，不能只考虑对机器人的惩罚，而是需要在另一个方向上对伤害加以防范。在上述思想实验的有限选项中，留给司法实践的余地不多。必须为人类驾驶者开通更高的驾驶权限，要么进行自我牺牲，要么承担事故责任。但如此一来，在没有考虑到人类的反应时间等因素的情况下，就会将人类驾驶者置于承担“道德运气”风险的无辜境地，对此设计者与生产商是应该背负责任的。争论是否应在设计和生产环节中取消自动驾驶汽车的方向盘，对解决惩罚缺漏问题毫无帮助。在本质上，此类建立“风险共同体”的操作会将矛盾转嫁给无辜群体。例如，骑手与平台算法共同承担因算法不合理导致的交通事故，会掩盖真正应该负责任的主体。

实际上，解释相关性决定了思想实验的有效性。例如，保罗·郎之万（Paul Langevin）试图用双生子佯谬在“动钟延缓”（time dilation）与相对性原理之间构造悖论，以解释理论上的“运动的相对性”与经验上的“动钟延缓的绝对性”之间的冲突，并得出对狭义相对论的批评，然而二者在解释上不相关，从而导致思想实验的无效。[①] 丹纳赫有意使用“机器人的高自主度”来解

① 参见孙圣：《科学思想实验的划界问题研究：技术细节及其缺省》，《自然辩证法研究》2022 年第 6 期，第 109~114 页。

释惩罚缺漏。不难发现，丹纳赫在用以说明惩罚缺漏的思想实验中设想的情景与他要批判的对象之间并不具有明确的解释相关性。对惩罚缺漏的争论而言，更合适的展开维度应该是拟人化而非自主度，前者属于伦理范畴，后者属于技术范畴，前者更具有解释相关性。丹纳赫在他的研究中有意地声明需要避免涉及拟人化争论，这就绕开了对惩罚缺漏的争论而言最为关键的维度。

三、机器人拟人化必须是客观的吗？

在笔者看来，拟人化与非拟人化之间的界限并不清晰，这才是惩罚缺漏的真正原因。以美国最高法院对救济性立法《美国伤残人士法》做出的扩张性解释为例，这一解释存在严重争议——当个人通过努力克服生理或智力限制时，对他们的法律保护反而不存在。[①] 实际上，在机器人时代，原本瘫痪的人类个体在植入机器人脊椎后，从伦理直觉上似乎有理由被视为具有完全责任能力的“风险共同体”。但这样一来，划定拟人化与非拟人化的界限将变得更加困难。

姑且不论让人类与机器人共同作为“风险共同体”来承担惩罚是否会导致丹纳赫意义上的“惩罚缺漏”，这样的立法伦理是存在哲学前提的。它取决于将机器人作为与人类相似的道德主体和惩罚对象是否成立。为此，我们需要考察机器人拟人化的真实来源。

机器人拟人化是否是客观的？“客观的”在这里的含义是能够作为一个外在的标准而被几乎全部人类接受。或者，至少在具体的案例中，在将特定的机器人作为考察对象时，人们能够使用相对客观的标准审视其拟人化与否吗？实际上，拟人化不仅可能源自人为的设定，正如前文论述的那样，还与具身工作环境相关。机器人拟人化难以是彻底客观的。

我们知道，机器人不仅可能伤人，还可能受到伤害。通常而言，人们谈论对财产的“损坏”“损毁”，而不使用“伤害”或“侵害”等词。将机器人视为被侵害的对象是否合适？为此，需要首先类比人类案件，其中对盗窃“非特定物”和“特定物”在惩罚上存在区别。

对机器人应被视为“特定物”还是“非特定物”的选择将在司法实践上决定人类施害者的犯罪性质，从而改变对人类施害者的惩罚程度。对此，相关的司法实践具有借鉴意义。例如，期待盗窃售价很高的特定艺术作品 A 的

① 参见［美］格林豪斯：《美国最高法院》，何帆译，译林出版社 2017 年版，第 27 页。

人类嫌疑人在入室后没有找到这件作品，而只找到了同样售价的作品 B，从而未能如愿，停止犯罪行为。这一行为在司法实践中通常被认定为盗窃未遂。相比之下，期待盗窃同样价值的 C 国货币的人类嫌疑人只找到了同样价值的 D 国货币，从而停止犯罪行为，则或可被认定为盗窃中止。同样地，对机器人相关案件的伦理反思不应将争论局限在人类财产权是否受到侵害之上，还应考虑机器人是否为“特定物”。

具体而言，关于是否可以将机器人视为被侵害对象的争论源于人们赋予机器人在设计意义之外的拟人化的特定情感。在施加侵害的人类看来，被侵害的机器人只具有财产属性，其与人类所有者的关系不具备财产要素之外的内容。然而，在被侵害机器人的所有者看来，机器人与自己之间具有财产归属之外的伦理关系。例如，养老机器人能够适应服务对象的个性化需求，不能被其他原厂设置的养老机器人替代，从而有了一定的家庭属性与伦理地位。进而，相较传统意义上的工具和机器而言，机器人所处的境遇或与猫、狗等家养动物更为接近，甚至超过猫与狗的伦理地位。

可见，拟人化作为情感标签，有时并非局限在物理上的外观设计或出厂设置方面，而是来自商业和实际生活场景，以及来自人为的赋予，产生于一种心理活动和主观行为。此类具有拟人化情感因素的“特定”个体的应用场景还有很多。例如，自我牺牲的拆弹机器人会被赋予“战友”等拟人化的情感标签。

然而，爆破机器人会使受攻击方的人类期待报复攻击方的人类，而非报复攻击方的爆破机器人。这一受攻击方的心理结果的出现，并不以爆破机器人是否被攻击方赋予了“战友”等拟人化伦理身份为前提，双方对惩罚行为是否足够以及对惩罚对象是否合理在认知上不存在匹配的可能性。这里的拟人化身份不是波普尔意义上的客观性的知识。

没有得到广泛情感认同的拟人化是非客观的，不具有可迁移性。在机器人伤人案件中，尽管对机器人一定程度的惩罚会对机器人所有者造成一定的情感上的损伤或产生震慑作用，但并不能强制认为，惩罚会使被机器人侵害的人类个体感到满意，惩罚也不能像丹纳赫相信的那样“实现人类的报复本能”。

由于拟人化不具有广泛的客观意义，故而单纯地惩罚机器人无法避免与机器人案件有关的惩罚缺漏问题，这也是前文论述的关于惩罚缺漏对象存在争议的深层原因。至少在现阶段，对人类有效的法律无法直接迁移以至于对机器人同样有效。问题在全新的意义上再次回到原点。在现阶段以及未来，人们如何从两个方向封堵惩罚缺漏问题？

四、以拟人度划分技术阶段何以可能

本文通过提出拟人度概念区分机器人技术的三个发展阶段，并针对不同发展阶段给出封堵惩罚缺漏问题的伦理和立法建议。其一，在较低拟人度阶段，将对传统商品有效的法律做直接迁移。其二，在中等拟人度阶段，将动物保护法和宠物管理有关条例做必要修改后进行迁移。其三，在较高拟人度阶段，将对人类有效的部分法律做修改后进行迁移。

具体而言，在根据拟人度给出不同技术阶段的判决标准时，会遭遇更多困难。我们知道，图灵测试（Turing test）将“与人类无法区分的智能”作为机器人是否实现拟人化的判决标准，但它早已过时。在特定领域中，人工智能相较人类是更强的，如新材料领域的 AlphaFold 和围棋领域的 AlphaGo，二者的成功均无须以还原人类思维为前提。

实际上，智能特征并非为人类特有，可以被人们精确地刻画并由机器载体实现。① 马文·明斯基（Marvin Minsky）认为，人工智能与人类智能在理解和学习的能力上都基于对困难问题的解决。② 尽管机器人未必能感同身受，但这并不妨碍机器人存在输出完全符合人类伦理标准的行为的可能性。同样地，未来基于对机器人惩罚相关的法律法规和伦理道德方面的知识，机器人可通过具身思想实验规划自身的行为。尽管未必能完全符合人类伦理标准，但机器人至少应该可以表现出它能理解其特定行为会导致怎样的人类伦理判断。这样，则较高拟人度的基础得以建立。

在笔者看来，较高拟人度建立在两个基础上：基础一，人类个体能够在全新的科技条件下有效理解有关机器人的伦理知识；基础二，机器人能够理解该伦理条件下的人机互动知识，并证明给全体人类，使其相信机器人能够在行为过程中展现对这些知识的理解。

可见，拟人度是双向的，既包括了机器人能否理解人类，也包括了人类个体能否相信理解是可能的。如果对机器人的惩罚能够体现这两个方向之间的对称性，则较容易实现对机器人惩罚缺漏的封堵。从概念上讲，双向的拟人度本身便可消解机器人惩罚缺漏问题。在具体的阶段判定上，人们还需要

① 参见 John McCarthy, Marvin L. Minsky and Nathaniel Rochester et al., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955, *AI Magazine*, Vol.27(4), 2006, pp. 12−14。

② 参见 Marvin Minsky, *The Society of Mind*, New York: Simon & Schuster, Inc., 1986。

在技术方案得以实现的同时，随时在理论上做必要的哲学反思，并在实践环节中训练人类接受这一现实。

对基础一与基础二的讨论需要在两个原则上达成基本共识：原则一，充分认识有关机器人的伦理问题具有的复杂性；原则二，对拟人度阶段的判定取决于机器人输出行为的伦理结果是否广泛地与人类个体的直觉相符合。基础一与基础二共同的关键是，通过对机器人行为伦理知识的理解，实现人们对机器人拟人度高低的判定。

然而，“机器人理解”与“理解机器人”存在大量的前提。拟人化的定义虽然属于伦理范畴，但拟人化的实现方式是技术主导的。从技术角度看，机器人拟人度是否存在上限？对这个问题的回应，将决定人们如何看待基础一。人类能否认可机器人行为是基于人类伦理知识的？如果确实存在拟人度上限，则或可导致机器人无法真正理解有关机器人的伦理知识，即不能实现基础二。

为了说明拟人度是否存在上限，需要考察拟人度的评价标准是客观的还是主观的。前文争论的问题与丹纳赫声称的“自主性”有一定的相关度，即机器人具有的主体性和目的性在层次上是否与人类的相当。由于现代科学理论的基本标准是“可测度”，[①] 故而如何测度和计算机器人的主体性与目的性成为问题的关键。这种测度是在空间规模上的，还是在时间尺度上的？笔者预计，未来对此问题的有效争论包括但不限于以下方面：光速限定了芯片的规模尺度和集成度，黑洞熵限定了智能体的信息存储密度，适应度函数限定了智能体的目的性层次在时间中的演化路径。[②] 这或将有助于预测技术社会的前景，以应对多元的机器人伦理困境。而现有的研究结论均难以证明机器人拟人度存在天然上限。

历史地看，对机器人智能最初的设计仅仅是辅助性的，而不具有主体性和目的性。因而，不能简单地将机器人具有的智能直接看做人类意识。其原因可能是由机器认识不透明性导致的限度等问题。[③] 机器人最终是否能生成等同于人类个体的意识尚存疑问。人们或许不能彻底相信基于完全不同的进

① 参见孙圣、万小龙：《测度对于观念的改变何以重要》，《哲学分析》2024 年第 6 期，第 138~150 页。

② 参见孙圣：《律则必然性对层级个体化问题的回应》，《自然辩证法通讯》2023 年第 5 期，第 42~48 页。

③ 参见董春雨：《从机器认识的不透明性看人工智能的本质及其限度》，《中国社会科学》2023 年第 5 期，第 148~166 页。

化路径和生存需求能够对同一伦理知识产生完全相同的理解。可见，基础一和基础二具有相关性，即基础二最终能否实现取决于基础一的判定。为此，需要首先从人类头脑与机器人在物理基础、硬件结构、知识来源等方面的相似性角度加以探讨。

从物理基础和硬件结构上看，智能系统都是物理符号系统（physical symbol system）。为满足理论模型，人类头脑将外部世界信息理想化地处理为便利属性，但偏离事实真相。① 在这一过程中损失的信息实际上是冗余（redundancy）。冗余描述了宇宙事件间的普遍联系。② 每个事件透过引力携带全息宇宙其他部分的信息。③ 包含冗余信息为湿件（wetware），不包含为干件（dryware）。冗余能解释外部世界连续变化与人类头脑离散观测之间的冲突。④ 以量子力学的退相干诠释为例，它将哥本哈根诠释的波函数坍缩解释为量子系统向环境泄露了冗余信息。⑤ 可见，人类头脑与机器人都不具有外部世界中隐含的丰富的冗余特征。

从知识内容的形态及来源看，机器知识在本质上也是由人类头脑加工后输入系统的。人类头脑与机器人的知识在形态上都具有干件的同源性。考虑到以 ChatGPT 为代表的初代人工智能都是基于注意力机制 ⑥ 的，世界向头脑中的映射从而被“拧干”。人类头脑无论在何种意义上难以理解机器知识，二者均源自涌现（emergence）机制，而非源自其他可能导致知识被“加湿”的新原理。机器人与人类头脑都具有干件特征，从而是相似的。

从立法伦理上讲，法律之所以能实现对人类个体侵害行为的惩罚，在于假定人类具有自由意志。⑦ 然而，此前关于机器人惩罚缺漏的争论完全建立在人类单方面推理上，未从机器人的具身认知角度加以考量，即没有考虑到

① 参见符征、李建会：《认知计算主义的六个里程碑》，《科学技术哲学研究》2015 年第 3 期，第 24 页。

② 参见 Lee Smolin, *Three Roads to Quantum Gravity*, New York: Basic Books, 2001, pp.53−54。

③ 参见 Roberto M. Unger and Lee Smolin, *The Singular Universe and the Reality of Time: A Proposal in Natural Philosophy*, Cambridge: Cambridge University Press, 2014, p.371。

④ 参见孙圣：《多世界解释的理论出发点及其哲学分析》，《自然辩证法研究》2023 年第 5 期，第 24~31 页。

⑤ 参见 Wojciech H. Zurek, Quantum Darwinism, *Nature Physics*, Vol.5(3), 2009, p.183。

⑥ 参见 Ashish Vaswani, Noam Shazeer and Niki Parmar et al., Attention Is All You Need, https://arxiv.org/pdf/1706.03762v7, pp. 1-15。

⑦ 参见孙圣：《思想的粒度与边界：泛化目的论的实现解释何以可能》，新华出版社 2020 年版，第 25 页。

机器人根据相应的法律环境和技术环境做出应有的具身规划，从而实现在一定程度上使自身行为符合其对伦理知识的理解。人们在对这一问题的现有理解中通常只从机器人的外在特征出发进行拟人化判定，而忽视考察机器人是否理解自身行为后果的伦理知识，容易让人产生额外的联想而引发误解。在惩罚尺度上的天然沟壑，其来源通常被认为是人类中心主义。而在笔者看来，它更多地是由以下问题导致的：人类与机器人目前尚不能对等地理解自身和对方行为的伦理后果。

无论争论结果如何，有一点都是确定的，即不必将复刻人类行为作为智能的唯一定义方式，可根据理性来抽象地定义智能。[①] 如果这个断言为真，那么机器人的智能可在不具备主体意识或目的性的情况下，实现对人类伦理知识的广义“理解”。然而，很多反对者声称，机器人与人类头脑在底层结构上的差异使机器人对人类伦理知识的理解存在上限。本文尝试从对拟人度未来的判决标准上讨论这一问题。

到目前为止，对任何一个科学理论的表述，如表达物理学原理的数学公式，都可以书写为由自然语言组成的、能够判决真值的命题形式。机器人对自身行为伦理后果的知识也可以被书写为命题形式。因此，只需证明自身理解了相应的命题，就可以声称对机器人行为后果的伦理知识实现了理解，从而实现了人类在整体上对机器人较高拟人度的普遍认可。

在人类头脑与机器人是否理解自身和对方行为后果的伦理知识的问题上，存在一定的检验方式。为避免透明性等问题对争论的影响，可考虑零知识证明（zero-knowledge proof）[②] 的方式以省去不必要的麻烦。对互不信任的通信双方，其中一方作为受试者，可在不泄露额外信息的前提下向作为检验者的另一方证明某个命题为真。

零知识证明在数理基础、算法和实现上涉及诸多领域的技术细节，但其定义并不困难——受试者声称的命题只有在为真时，才会得到检验者的确认。这样，基础二成立的关键就转化为：假设存在一个人类专家 A，他声称一个命题 P“A 自己能理解关于机器人特定行为是否符合人类伦理判断的知识 T”，其他人类个体 B 能否对命题真值加以判断。如果能判断命题 P 为真，则知识

① 参见 Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, London: Pearson, 2021, p. 1。

② 参见 Shafi Goldwasser, Silvio Micali and Charles Rackoff, The Knowledge Complexity of Interactive Proof Systems, *SIAM Journal on Computing*, Vol.18(1), 1989, pp.186–208。

T 是可理解的，并且可被人类个体理解。

不难发现，对 A 声称的命题 P 的真值的判断，并不建立在施加判断的 B 对这个知识 T 是否理解的前提之下。如果这个伦理知识 T 是关于机器人行为后果的，并且命题 P 的真值是可由 B 判断的，则 T 是可理解的。声称命题 P 的主体 A 可以是人类个体，也可以是机器人个体。如果存在一个由机器人行为后果导致的伦理知识，那么可由机器人个体声称其实现了基础二，再由人类个体对基础一做出为真的判断，如此则实现了零知识证明，进而，没有理由认为机器人不具有较高的拟人度。

无论如何，技术的发展可能会出现“路径依赖”（path dependency），即对技术发展路线的选择会影响技术的最终状态以及社会形态。机器人伦理研究的干预或多或少会对技术发展路线的选择以及社会演化的方向产生正面或负面的影响。如果确实如人们担心的那样，人类社会遇到下一个阶段的各类技术瓶颈，如意识心灵的难问题、结果的不可预测、过程的不透明性、知识的不可理解性、范式的不可迁移性等，那么可能会出现技术社会发展的停滞，包括暂时的和永久的两种类型。如果出现类似于永久停滞的情况，那么由于路径依赖而停滞的技术状态将在相当长的一段时间内决定人类的福祉。在公平与效率之间，人们必须做出选择。人类已经步入人机共存社会的第一阶段，机器人伦理研究具有十分紧迫的历史与现实意义。

五、结语

丹纳赫理论建立在机器人自主度的维度上，而对机器人惩罚缺漏而言，真正有效的争论应建立在拟人化的维度上。过于强调对机器人的惩罚缺漏将导致惩罚过度等不良后果。相反，惩罚缺漏的真正来源在于没有考虑到将人作为惩罚对象的情形。其深层原因在于，惩罚依赖于拟人化的程度。在现有的技术环境下，人们难以保证拟人化的客观性，界限不清导致惩罚难以使各方满意。这从根本上拒斥了将惩罚缺漏作为机器人伦理模型的判决标准。

本文提出了人类头脑与机器人双向拟人度标准，用以对未来人机共存社会发展阶段进行技术性的划分，并对多种潜在的反对声音给予了回应。以处理外部世界冗余问题的视角，论述了人类头脑与机器人在物理基础与硬件结构、目的性与主体性的层次、伦理知识形式的来源等方面的相似性。从零知识证明的角度，初步设想了对图灵测试的超越何以可能。

·笔谈·

机器人伦理学前沿问题

刘永谋等

【主持人语】经过几十年的发展，机器人技术已达到某种“临界点”，表现为人形机器人（humanoid robot）技术加速发展。工业和信息化部在2023年10月印发的《人形机器人创新发展指导意见》开篇即高屋建瓴地判断，人形机器人“有望成为继计算机、智能手机、新能源汽车后的颠覆性产品”。人形机器人的科技研发和应用同样面临科技风险和科技伦理问题，受到全社会的广泛关注。由此，机器人伦理学成为近年来的热门研究领域，尤其是应用伦理学重要的“理论增长点”。本次笔谈由6篇文章组成，聚焦机器人伦理学发展前沿，抛砖引玉，以期推动该领域研究的进一步发展。刘永谋和白英慧讨论拟人论意识形态在机器人伦理学建构中的基础作用，分析机器人拟人论在该领域流行的原因及启示。刘鹏考察了机器人伦理风险的发生机制及治理原则，基于此指出人类社会应通过与机器人的互动、互构，构建一种新型的人机有机结构。程林在跨文化视域下考察了机器人拟人化及恐惑现象，对中国人机协存观念和机器人设计理念提出了建议。谭笑考察了小数据主义技术路线在隐私保护和权力结构均衡等方面的优势，论证了由于所需知识类型不同，这一路线不太适用于社交机器人领域。孙圣引入拟人化分析，阐述了机器人伦理模型的惩罚缺漏不可避免，以此反驳直接将之作为判决标准的可行性，提出拟人度概念，用以划分人机共存社会发展阶段，并论证了划分何以可能。杨庆峰和朱清君考察了人形机器人导致的社会角色的临时替代和永久替代问题，并探讨了由人形机器人角色替代带来的伦理学挑战。

（刘永谋，中国人民大学哲学院教授、博士生导师）

【关 键 词】机器人　人形机器人　伦理学　机器人伦理学

【作者简介】刘永谋，中国人民大学哲学院教授、博士生导师；白英慧，中国人民大学哲学院博士研究生。刘鹏，南京大学哲学学院教授、博士生导师。程林，广东外语外贸大学外国文学文化研究院教授、阐释学研究院兼职研究员。谭笑，首都师范大学政法学院教授。孙圣，西北师范大学哲学与社会学院副教授、硕士生导师。杨庆峰，复旦大学科技伦理与人类未来研究院教授；朱清君，复旦大学社会发展与公共政策学院博士研究生。

【中图分类号】B829　【文献标识码】A

【文章编号】2097−1125（2025）06−0005−55

机器人伦理学的拟人论基础*

刘永谋　白英慧

近年来，人工智能、传感器、机器人控制与动力学、云计算与物联网、人工肌肉与柔性材料等技术的不断突破，使机器人更加灵活与自主，并使其应用场景愈加多元化、精细化。同时，机器人的研究、设计、制造和使用面临人类失业、隐私泄露、情感欺骗、责任分配等十分棘手的伦理问题，这些问题对机器人技术的发展及人类社会生活产生了不可忽视的重大影响，必须结合具体情境加以认真研究。在此背景下，机器人伦理学（roboethics）兴起并持续火热。顾名思义，机器人伦理学是研究有关机器人的伦理问题的学问。凯斯·阿布尼（Keith Abney）总结了机器人伦理学研究对象的三层含义：第一，机器人技术专家的职业道德；第二，为自动化机器人编写的道德规范的代码程序，即机器人自己的而非人类的准则；第三，机器人在具备进行伦理推理的自我意识能力时自行选择的伦理准则。① 换言之，机器人伦理学研究

* 本文系国家社会科学基金重大项目“现代技术治理理论问题研究”（21&ZD064）的阶段性成果。

① 参见［美］帕特里克·林、凯斯·阿布尼、乔治·A. 贝基主编：《机器人伦理学》，薛少华、仵婷译，人民邮电出版社 2021 年版，第 35 页。

Abstracts

Frontier Issues in Roboethics
Liu Yongmou et al.

【Abstract】 After decades of development, robotic technology has reached a certain "critical point", manifested in the accelerated development of humanoid robot technology. The Ministry of Industry and Information Technology issued the *Guidelines for the Innovation and Development of Humanoid Robots* in October 2023, which made a high-level judgment at the outset that humanoid robots "are expected to become a disruptive product after computers, smartphones and new energy vehicles". However, the research, development, and application of humanoid robots also face scientific and technological risks as well as scientific and technological ethics issues, drawing widespread attention from the whole society. As a result, robot ethics has emerged as a hot research field in recent years, particularly as a significant "theoretical growth point" within applied ethics. This special issue features six invited articles that focus on the cutting-edge developments in roboethics, aiming to spark further discussion and advance research in this field. Liu Yongmou and Bai Yinghui discuss the fundamental role of anthropomorphic ideology in the construction of roboethics, and analyze the reasons for the prevalence of anthropomorphism in this field and its implications. Liu Peng examines the mechanisms of the occurrence of robot ethical risks and the principles of governance. Based on this, he points out that human society should build a new type of human-machine organic structure through interaction and mutual construction with robots. Cheng Lin explores the anthropomorphism and uncanny valley phenomenon of robots from a cross-cultural perspective, offering suggestions for the Chinese-style human-machine co-existence and robot design. Tan Xiao examines the advantages of the small data technology roadmap in terms of privacy protection and power structure balance, and demonstrates that this roadmap is not very suitable for the field of social robots due to the different types of knowledge required. Sun Sheng introduces anthropomorphism analysis to demonstrate the inevitability of the retribution gaps of ethical models of robots, thereby refuting

the feasibility of using them directly as a criterion for judgement. He proposes the concept of degree of anthropomorphism as the demarcation of development stages of human-machine coexisting societies and justifies its plausibility. Yang Qingfeng and Zhu Qingjun investigate the temporary and permanent substitution of social roles caused by humanoid robots, exploring the ethical challenges posed by the role substitution of humanoid robots.

(Liu Yongmou, Professor and PhD Superviser, School of Philosophy, Renmin University of China)

【Keywords】robot; humanoid robot; ethics; roboethics

Promoting the Common Values of Mankind: China's Contribution to the Advancement of International Rule of Law

Li Lin

【Abstract】The proposal of new ideas regarding the common values of humanity has not only provided a new value foundation for people of all countries to join hands in building a community with a shared future for mankind, but also offered strong value guidance for China's participation in promoting theoretical, institutional and practical innovations in the international rule of law. Since the founding of the People's Republic of China, particularly since the 18th National Congress of the Communist Party of China, China has upheld the banner of human values and civilizations and made significant contributions to safeguarding world peace and promoting the development of the international rule of law. Hence, as a responsible major power emerging on the global stage and engaging in international affairs, China must adeptly employ the rule of law. In order to employ the rule-of-law thinking and rule-of-law based approach to boost the building of a community with a shared future for mankind, greater emphasis should be placed on the coordinated advancement of domestic and foreign-related rule of law. China is also required to promote the common values of humanity, advance foreign-related rule of law initiatives, and actively engage in the development of the international rule of law, thereby contributing more Chinese wisdom and strength to the progress of international rule of law.

【Keywords】common values of humanity; domestic rule of law; foreign-related rule of law; international rule of law; rule of law civilization